Published in: Barbu Mititelu, Verginica/Pekar, Viktor/Barbu, Eduard (ed.): Proceedings of the Workshop "Semantic Relations. Theory and Applications", 18 May 2010, at the International Conference on Language Resources and Evaluation (LREC) 2010, Malta. - European Language Resources Association (ELRA), 2010. p. 37-46.

Consistency of Sense Relations in a Lexicographic Context

Peter Meyer, Carolin Müller-Spitzer
Institut für Deutsche Sprache (IDS)
Mannheim, Germany
E-mail: meyer@ids-mannheim.de, mueller-spitzer@ids-mannheim.de

Abstract

The representation of semantic relations between word senses of different entries in a dictionary is subject to a number of consistency requirements. This paper discusses the issue of maintaining and accessing consistent information on cross-references between sense-related items in electronic dictionaries from a mainly text-technological point of view. We present a number of consistency criteria for cross-referencing related senses and propose a practical approach to handling sense relations in an online dictionary. Our proposal is currently being tested in a large ongoing online dictionary project for German called *elexiko*. We focus on three different aspects of the dictionary development and editing process where consistency is an important issue: lexicographic data modelling, implementation of a lexicographic database system for an electronic dictionary, and development of practical tools for the lexicographer˝s workbench.

1. Introduction

Semantic relations between lexicographic items, such as synonymy and hyponymy between specific senses of different lexemes, are typically encoded as cross-references in the respective entries in a dictionary.

The necessity of keeping the reference structure of a dictionary *consistent* raises a number of conceptual and practical issues. In the context of describing lexical-semantic relations in dictionaries, consistency may require that, among other things, bidirectional relations, as existing in paradigmatic sense relations, are given for both reference points between which a specific relation holds. For example, if *require* is given as a synonym in the entry *demand*, then *demand* should also be listed as a synonym in the entry *require*. This is a form of consistency that is important for the underlying lexicographic data model as well as for the dictionary user.

As a matter of fact, however, consistency in bidirectional references is rarely met. In Figure 1, three entries taken from Duden: "Das Synonymwörterbuch" (2007), a conventional German dictionary of synonyms, are shown: *arbeitsunfähig* (unfit or unable to work), *dienstunfähig* (disabled, unfit for service), and *erwerbsunfähig* (unable to work, incapacitated). The meaning descriptions of these three entries are semantically very close. The terms constitute a set or cluster of synonyms. Nevertheless, there are striking inconsistencies. For example, in the entry *arbeitsunfähig*, the synonym *erwerbsunfähig* is missing although *arbeitsunfähig* is given as a synonym of the head word *erwerbsunfähig*. In addition, *dienstunfähig* is not listed as a meaning equivalent to *arbeitsunfähig*, whereas in the entry *dienstunfähig,* both *arbeitsunfähig* and *erwerbsunfähig* are listed as synonyms (cf. Müller-Spitzer 2010).

arbeitsunfähig
bettlägerig, krank, unpässlich; (*bildungsspr.*): indisponiert; (*oft emotional*): malade.

dienstunfähig
arbeitsunfähig, erwerbsunfähig, invalide, krank, nicht arbeitsfähig, nicht dienstfähig, nicht einsatzfähig, untauglich.

erwerbsunfähig
arbeitsunfähig, behindert, dienstunfähig, invalide; (*Amtsspr.*): schwerbehindert, schwerbeschädigt.

Figure 1: Entries *arbeitsunfähig*, *dienstunfähig*, and *erwerbsunfähig* from Duden: "Das Synonymwörterbuch" (2007).

It could be argued that consistency is not of particular importance here. Presumably most lexicographers attempting to compile a reference dictionary of synonyms chiefly aim to provide an abundance of words with similar meanings that can be substituted for each other: Their intention is not to depict theoretical lexical-semantic structures as lexicographic information, cf. also (Lew, 2007). However, it is argued here that, as the entry *arbeitsunfähig* in particular illustrates, a more consistent approach would help to provide the dictionary user with better information. Presumably any lexicographer would have added *erwerbsunfähig* as a synonym of *arbeitsunfähig* to this dictionary, if the incomplete listing had been noticed.

More generally, consistency of cross-references means that, depending on the overall design and purpose of the dictionary, its reference structure should reflect certain formal properties of the underlying lexical and semantic structure. A simple example of such a property is a symmetry constraint on synonymy: If word sense *S1* of lexeme *L1* is synonymous with word sense *S2* of lexeme *L2*, then, trivially, *S2* is also synonymous with *S1*. This

implies, as we have seen above, a possible corresponding requirement on the cross-reference structure of a dictionary: In many lexicographic contexts, if the section on *S1* in the entry for *L1* contains a synonymy reference to the section on *S2* in the entry for *L2*, then there should be a corresponding reverse reference in the *L2* entry. Similarly, if *S1* stands in a hyponymy relation to *S2*, then *S2* is a hypernym of *S1*. In this case, however, enforcing the corresponding possible requirement on reference structure is not feasible in conventional print dictionaries since this would imply that each and every hyponym of a lexeme must be included in its entry. But, as already noted above, not even the symmetry of the synonymy relation is usually enforced in standard dictionaries, cf. also (Müller-Spitzer, 2007).

Compared to print dictionaries, users of electronic dictionaries are much more likely to be confused by missing reverse links for a synonymy reference to another article because following links to sense-related items in an electronic dictionary is faster and more straightforward than looking them up by leafing through a printed dictionary. If a synonym is given for a specific sense in an entry and in the link-targeted entry this headword is not mentioned as a synonym, users are probably surprised by the lack of reverse linking. Here, a formal inconsistency at the level of data modelling easily leads to an inconsistency (in a less formal sense of the word) on the level of presentation and, hence, in user experience. Moreover, keeping track of all semantic relations represented in a lexicographic database is an elementary and essential prerequisite for lexicographic work on an electronic dictionary. It would be very useful if lexicographers were automatically informed that the entry is already mentioned as a target in another entry when they start to write a dictionary entry. Protecting dictionary authors from producing inconsistencies this way calls for extensive computer assistance, particularly when large amounts of data are involved.

On a terminological note, we will say that in both the synonymy and the hyponymy case two *unidirectional* references may stand in a *reverse relation* to each other and then together form a *bidirectional* reference. Provided that the unidirectional components of a bidirectional reference are stored in separate places, they must *correspond* to each other in that they (a) encode reverse semantic relations and (b) the target item of one unidirectional reference is the source item of the other and vice versa. This will be called the *correspondence requirement* for bidirectional links. Obviously, this is a different kind of consistency since the correspondence requirement for an actually bidirectional synonymy reference must be satisfied regardless of the question whether *all* synonymy references should be bidirectional.

2. XML modelling of Sense-Relation References: The Case of *elexiko*

We will discuss conceptual and implementational aspects of maintaining and controlling referential consistency in a concrete case, namely, the German corpus-based monolingual online dictionary *elexiko* that is accessible free of charge under *www.elexiko.de* and forms part of a long-standing and ongoing research project of the Institut für Deutsche Sprache (Institute for the German language), cf. (Haß, 2005), (Klosa et al., 2006). *elexiko* is still in progress (elexiko, 2003 seqq.); thus, this dictionary is not a complete reference book following an alphabetical compiling procedure.[1]

The lexicographic data pertaining to each *elexiko* entry are realised as a single XML document. All documents conform to a highly granular structural layout as defined in a complex XML Document Type Definition (DTD). The structural layout is strictly based on lexicographic content; any presentational aspects, such as typographic details, are taken care of by XSL transformations that generate HTML documents from the XML data.

In order to demonstrate the internal makeup of *elexiko* entry documents, we present a fragment of a typical XML representation. To ease comprehension, we will not use the original element names used for *elexiko* documents, but some hopefully self-explanatory English equivalents. The XML structure presented here is slightly simplified where this does not affect the topic under discussion. Boldface type is used to indicate data that is used to uniquely specify a particular reference to a sense-related item.

```
<elexiko-article id="1234">
 <general>
  <lemma-sign>Familie</lemma-sign>
 </general>
 <sense id="relatives">
  <usage>
   <paraphrase>
    Mit Familie bezeichnet man eine Gruppe von
Personen, die durch Geburt oder durch Heirat
miteinander verwandt sind. In engerem Sinn
bezieht sich der Sprecher mit Familie auf eine
Lebensgemeinschaft, die aus Eltern und
Kindern besteht, in weiterem Sinn auch auf
eine Gemeinschaft, die mehrere Generationen
umfasst und zu der z. B. die Großeltern, die
Geschwister der Eltern und Großeltern ein-
schließlich deren Angehörige usw. gezählt
werden.
   </paraphrase>
   <paradigmatic-relations>
    <partonymy>
     <item  articleID="9999"
            senseID="female descendant"
            subsenseID="0">
       Tochter
```

[1] In this paper, we will not discuss the linguistic and lexicographic foundations for the kind of XML modelling and for the treatment of sense relations in *elexiko*. However, there is ample literature that relates *elexiko* to other approaches in electronic lexicography, cf. Storjohann 2009 and 2010 and www.owid.de/elexiko_/pgProjektveroeffentlichungen.html resp. http://www.owid.de/elexiko_/pgVortraege.html for up-to-date references.

```
      </item>
      <item  articleID="3737"
             senseID="mother and father"
             subsenseID="0">
        Eltern
      </item>
     </partonymy>
    </paradigmatic-relations>
    <subsense id="dynasty">
     <subsense-paraphrase>
 Mit Familie bezeichnet man eine angesehene,
wohlhabende, einflussreiche bzw. adlige
Personengruppe, deren Mitglieder durch Geburt
oder Heirat miteinander verwandt sind.
     </subsense-paraphrase>
     <paradigmatic-relations>
      <synonymy>
       <item articleID="5678"
             senseID="dynasty"
             subsenseID="0">
        Haus
       </item>
       <item articleID="1066"
             senseID="dynasty"
             subsenseID="0">
        Dynastie
       </item>
      </synonymy>
     </paradigmatic-relations>
    </subsense>
   </usage>
  </sense>
  <sense>
  </sense id="biological taxon">
   …
</elexiko-article>
```

The root element of each entry document has an attribute @id, its *article ID* – a string representation of an integer number uniquely identifying the entry. It contains one <general> element with sense-independent information (relating to, e.g., orthography and morphology) and arbitrarily many <sense> elements representing different word senses. No distinction is made between polysemy and homonymy.

The lemma sign [for terminology cf. (Hausmann & Wiegand, 1989)] is part of the general, that is, sense-independent information in the article as specified within the <general> element. In our sample entry with an article ID of "1234", the German equivalent to „family" has the citation form (nominative singular) *Familie*.

Each word sense is represented by a <sense> element with an attribute (a *sense ID*) that identifies this sense uniquely within the article. The most salient word sense of *Familie* might be paraphrased as „group of close relatives of a person". Using English IDs for the purpose of this article, we might choose "relatives" as the ID. The ID is not supposed to be a concise hint at the semantics of a sense; it just serves as a convenient mnemonic. In the XML document, a short explanation of the contexts associated with the word sense "relatives" is stored in a <paraphrase> element. For illustration purposes, a second word sense of *Familie* used in biology is shown in the XML fragment above.

The word sense with the ID "relatives" is assumed to have a specialized *subsense* in German, namely, „group of relatives who play an important role in society". This subsense appears nested inside the appropriate <sense> element as a <subsense> element with a *subsense ID* attribute "dynasty".

Figure 2 shows a partial view of the *elexiko* entry on *Familie* as it is presented to the user in a web browser. The sense and subsense IDs – here in their original German appearance, for example, "Verwandte" for "relatives" and "Dynastie" for "dynasty" – serve as headings for the different senses and subsenses. In this particular view, the meaning explanations as stored in the <(sub-sense-)paraphrase> elements are given.

Figure 2. A part of the HTML-based online presentation of the entry *Familie* in *elexiko* (screenshot).

In *elexiko*, references to sense-related items, henceforth *paradigmatic references*, always relate specific word senses of two entries. The "dynasty" subsense of *Familie* contains a synonymy reference to the corresponding sense of the entry *Haus* („house"). The type of sense relation (named *paradigmatic relation type* here) is encoded as a <synonymy> element inside the <subsense> element. A word (sub)sense may have more than one synonym, so each synonymous word sense is to be given as a separate <item> element enclosed by <synonymy>. Our sample fragment lists another synonym for the same word sense, namely, *Dynastie*. In addition, two partonyms for the "relatives" sense of Familie are given, namely: *Tochter* („daughter") and *Eltern* („parents"). The attributes and text content of each <item> element provide a complete specification of the end point or *target address* of the reference, that is, the lemma sign and article ID of the entry *Haus* as well as the sense and subsense IDs of the word sense referred to. If the target address concerns a word sense but not a subsense, "0" is used as subsense ID.

To sum up, three observations may be made at this point. First, in *elexiko*, three strings are used to uniquely identify the target address in a reference to a related item, namely,

- the ID of the target entry as a whole;

- the ID of the target sense; and, where applicable,
- the ID of the target subsense.

Second, all lexicographic information on sense relations targetting (sub)senses of other dictionary articles is stored in the individual entries‟ subsection (XML element) pertaining to the source address. Specifically, the information on the three source IDs as well as on the type of sense relation is distributed among different nested ancestor elements of the element containing the target specification. Third, outgoing references are stored in a strictly local fashion, that is, only in the source article. There is no indication in the source XML document as to whether a consistent reverse reference exists in the corresponding place in the target article.

In order to obtain all necessary information on a particular reference, the XML document containing the reference must be parsed, which is an expensive operation in terms of database operations. Ensuring consistency of cross-reference information in one document *D* inevitably requires parsing all documents referred to in *D* as well as all documents referring to *D*.

3. Criteria for Consistency in Cross-Referencing Sense-Related Items

From a lexicographer‟s point of view, there are many different ways in which a cross-reference might fail or be inconsistent. This section enumerates important criteria for evaluating the consistency of paradigmatic references in *elexiko*. As was said before, we concentrate on aspects of the data structure, not on general lexicological-lexicographic considerations.

In a well-formed and valid XML document representing an entry in *elexiko*, any paradigmatic reference must meet the following requirements:

- It must be **complete**: All necessary pieces of information related to a particular type of reference must be present. To a certain degree, completeness can be enforced through an appropriately specified DTD or XML schema. However, a common and unavoidable problem in the process of compiling a dictionary is the need to make preliminary incomplete references to target addresses that do not exist yet. In *elexiko*, paradigmatic references to a word sense in an entry not yet edited use the dummy word sense ID “0”.
- It must be **well-formed**: All required pieces of information must conform to formal specifications. Again, certain aspects of well-formedness cannot be captured by means of an XML schema, for example, conventions regarding allowed formats for different ID types.
- It must be **valid**, that is, it must point to an address that really exists in the lexicographic database. Note that validity presupposes both well-formedness and completeness of the reference. A particular prerequisite for validity is factual **consistency**: Different parts of a paradigmatic reference must not contradict each other. For instance in *elexiko*, a target address contains both the ID of the target entry and its lemma sign. Of course, the lemma sign specified in the target entry with that ID itself must be identical to the one given in the reference.

Let us call a reference that fulfills all of the above criteria a *correctly specified* unidirectional reference. Correctly specified references might still be lexicographically *inadequate* in relating wrong addresses or picking the wrong paradigmatic relation. Lexicographic adequacy cannot be checked by an automated procedure and constitutes yet another, very important kind of consistency requirement.

For a given unidirectional reference *R*, additional conditions are needed in order to define whether another unidirectional reference *R'* counts as a *potential reverse reference*, such that *R* and *R'* together form a bidirectional reference. The necessary requirements may be stated as follows:

- The type of paradigmatic relation must be a candidate for a bidirectional reference.
- Both *R* and *R'* must be correctly specified. Special provision must be made for the case that one or both of the references are specified correctly only on the dictionary entry level, but are not (yet) *complete* on the word sense level. This situation typically arises when the target article has not been edited yet.
- The correspondence requirement stated above must be met. If one of the references is not yet complete, this requirement must be relaxed to state that the target address of the incomplete reference must either be identical to the source address of the reverse reference or refers to a larger part of the entry that contains this source address.

If there are no potential reverse references for a given paradigmatic reference *R*, this might count as an instance of inconsistency in case bidirectionality is specified to be compulsory for the given paradigmatic relation. For example, *elexiko* employs a very narrow lexicographic concept of synonymy for which compulsory reciprocity is indeed a sound requirement.[2] If potential reverse references can be found in the lexicographic database, different cases may be distinguished according to which of these references are completely specified and whether there is more than one candidate reverse relation in the database. In case both unidirectional references *R* and *R'* are correctly specified and fulfill the correspondence requirement, we may classify the resulting bidirectional reference as correctly specified. Again, a correctly specified bidirectional reference might still be lexicographically inadequate.

This brief overview should suffice to demonstrate some of the intricacies of managing consistency issues in dictionaries. These problems must be dealt with at several

[2] In this paper, we simply use synonymy as a typical example candidate for a symmetric sense relation. Actual decisions on how to model sense relations will depend on the lexicographic setting and are independent of the conceptual and implementational points of the paper; hence, our approach can just as easily be applied to other sense relations such as antonymy: *elexiko* distinguishes between five categories of antonymy several of which are candidates for compulsory reciprocity.

stages of the process of conceiving, implementing, and editing dictionaries. The following sections will examine some of these stages in turn and discuss the merits and pitfalls of possible solutions.

4. Making Dictionary Entries Consistent: Considerations on Data Modelling

At first glance, a conceptually clear and simple solution to inconsistency threats in a lexicographic database seems to commend itself: Detach all reference-related information from the entry documents and put it in a separate table. After all, such a table (which we will call a *reference table* for short) would be the standard solution for modelling many-to-many relationships in a relational database. Each row in a reference table corresponds to a unidirectional or bidirectional paradigmatic reference. The columns specify the paradigmatic relation type and the three ID strings of source and target adress. The relational table might just as well be represented in an XML format. A sample entry for a *unidirectional* paradigmatic reference could then roughly look like this (cf. Section 2):

```
<reference relation="synonymy">
  <srcLemmaSign>Familie</srcLemmaSign>
  <srcEntryID>1234</srcEntryID>
  <srcSenseID>relatives</srcSenseID>
  <srcSubSenseID>dynasty</trgSubSenseID>
  <trgLemmaSign>Haus</srcLemmaSign>
  <trgEntryID>5678</trgEntryID>
  <trgSenseID>dynasty</trgSenseID>
  <trgSubSenseID>0</trgSubSenseID>
</reference>
```

In a similar XML representation, compulsory *bidirectional* references can be coded in a redundancy-free way that compliance with the correspondence requirement is guaranteed:

```
<reference relation="synonymy">
  <entry>
    <lemmaSign>Familie</lemmaSign>
    <entryID>1234</entryID>
    <senseID>relatives</senseID>
    <subSenseID>dynasty</subSenseID>
  </entry>
  <entry>
    <entryID>5678</entryID>
    <senseID>dynasty</senseID>
    <subSenseID>0</subSenseID>
  </entry>
</reference>
```

In ontology-based systems, this approach might be a sensible choice for modelling sets of synonymous senses since consistency is enforced when each *set* of *n* word senses is indeed represented as a *set* of XML elements instead of a group of $n(n\text{-}1)$ separate unidirectional references. Still, non-overlap of different sets of synonyms cannot be enforced this way. Aside from that, all entry and sense IDs in a reference table entry must themselves be correctly specified. This constitutes yet another consistency problem.[3]

As soon as other kinds of sense relations have to be considered for the data model, such as paradigmatic relations that are only *potentially* bidirectional, the disadvantages of a separate reference table will, in most cases, outweigh the benefits.

To begin with, a serious drawback of a separate data model for reference-related information becomes apparent when *entry-specific* information on paradigmatic relations is to be provided. In *elexiko*, for instance, sense-related items belonging to a given word sense in an entry are ordered according to corpus salience and discourse relevance. In such situations, the individual entries would have to include references to locations in the reference table, which would mean replacing one consistency issue with another. This problem is an indication for a more general need to separate two concerns, that is, to provide lemma-specific and lexicographically relevant information on sense relations on the one hand and to infer or keep track of all existing sense-relations between dictionary items on the other.

Introducing a separate reference table considerably complicates the editing process for dictionary entries since two tables must be modified concurrently and kept in agreement. As a consequence, manually editing the XML representation of an article becomes virtually impossible because it is too confusing and error-prone. A separate software tool would be needed just to keep the two database tables in synch at any time and to present all relevant entry-related reference table information in a perspicuous way to the lexicographer. As a final point, deciding which *types* of cross-references to pull out into a reference table and which to leave in the entries can be a delicate decision that cannot easily be changed later.

Everything considered, we believe that in most cases a minor improvement in handling compulsory bidirectionality will not justify the numerous administrative and conceptual complications induced by the introduction of a separate reference table. As a consequence, we strongly favour a maximally parsimonious data model for electronic dictionaries that leaves all reference-related information strictly within the respective entries.

5. Handling References on the Implementation Level

For *elexiko*, the „local" alternative outlined above has been opted for so that all unidirectional references are encoded solely within the respective entry documents and no separate data structure for bidirectional links is needed. This

[3] Partitioning all word senses of a language into equivalence classes presupposes transitivity of the synonymy relation. However, a range of philosophical, semantic, lexicological, and lexicographic arguments against the transitivity of synonymy have been advanced. Quine's insistence on the context-specific nature of synonymy springs to mind, cf. (Bosch, 1979) for a succinct overview. See (Storjohann, 2006) for a range of lexicological and lexicographic observations on synonymy that bear on this important issue.

means that, at least in principle, all management and information access tasks concerning (paradigmatic) references could be processed through queries on the XML representations of the dictionary entries. However, performance considerations regarding the underlying database system suggest a different strategy. As noted above, checking for inconsistencies in an entry's references would entail (a) searching the database for XML documents that contain certain information, that is, references to a given entry and (b) parsing these XML documents. Compared to a standard search operation in a relational database table, searching through hundreds of thousands of complexly structured XML documents is already a very expensive database operation, in terms of both time and CPU load, even if highly optimized indices (cf. Müller-Spitzer & Schneider, 2009) are used. Parsing the relevant XML documents is even more costly, no matter whether the parsing is done in the database system itself or on a client system.

As long as merely individual entries are checked for reference inconsistencies by a lexicographic tool (see next section), the necessary searching and parsing processes on the XML instances in the Oracle-based *elexiko* system take a few seconds at most. More demanding tasks such as the following ones are out of the question without a separate handling of reference information:

- searching all dictionary entries for inconsistent references, paradigmatic or other;
- processing complex queries requiring a recursive traversal of a possibly large number of referential links, such as for
    - visualising link trees starting from a given word sense;
    - finding minimal link paths between addresses;
- enabling end users of the dictionary to formulate and process complex queries on referential structure.

However, a simple and effective solution to the performance bottleneck of XML processing is available: One can simply *copy* all information pertinent to paradigmatic references to a separate relational database table. Afterwards, complex queries on cross-reference structure can be processed on this relational table using fast standard SQL queries. Initial construction of the additional table – which will be called *link table* in this paper only to distinguish it terminologically from a *reference table* as defined above – can be accomplished using a rather simple XQuery construct. This can be a time-consuming operation, but it needs to be done only once. Afterwards, the link table must automatically be updated each time an entry is altered, added, or deleted. To this end, a so-called trigger is installed in the database. The trigger starts a stored update procedure on the link table whenever the main table that contains the XML documents undergoes a change.

A link table may have exactly the same structure as a reference table. The difference to notice is that a link table does not contain any new information over and above the table of dictionary entries; it simply mirrors reference-related aspects of the dictionary entries. In other words, a link table is not part of the data model.

Even though the link table does not contain any information that is not already present in the XML instances, it offers several distinct advantages. It abstracts from the particularities of representing information in the XML format of the entries; specifically, as noted in Section 2, source and target of a reference are necessarily encoded in completely different ways within the entries while they can be represented in a simple and uniform format in the link table. Accessing the link table does not require parsing: it only requires standard relational database queries. Even if the information is represented as XML, modern database systems can transparently map it to an underlying relational representation, rewriting XPath expressions as SQL queries. In the Oracle database system used for *elexiko*, this is called "XML/SQL duality". Even though exact figures depend on a wide variety of factors, information extraction from an underlyingly relational link table may very well be 100 times faster than parsing dictionary entries. Oracle uses a dimension-less quantity named *cost* to measure the database system load for a query; and indeed, in terms of cost, looking up and parsing complex XML-based entries access might easily be more than 1000 times more expensive than a link table query.

Overall, modelling references in a strictly „local" fashion as an integrated part of the pertinent source entry is an approach both theoretically sound and pragmatically viable. Database performance can be enhanced dramatically through the use of a relational link table that provides fast access to the reference structure. The solution is robust in that it does not necessitate additional software tools for the editing process or a refactoring of existing database resources. The question which cross-reference relations should be included in the link table does not amount to a vital decision that is difficult to change afterwards.

A further decision has to be made as to whether bidirectional links should be encoded as two different and independent unidirectional entries (table rows) in the link table or rather be handled in a separate and possibly less redundant way, for example, as shown in the second XML example of Section 4. However, there are reasons to prefer the more redundant representation. In a typical setting where the database system has to process large numbers of potentially complex user queries on cross-reference structure, the time penalty induced by having to look up one more table row for a consistency check hardly matters. On the other hand, editing links in the entries has more complicated reverberations for a link table with a separate storage format for bidirectional links. If, for instance, a newly added article contains a paradigmatic reference that is reverse to an already existing one, the latter has to be deleted from the table while a new bidirectional reference is added.

6. Aiding the Lexicographer: Tools for Safeguarding Consistency

The implementation aspects that we focused on in the previous section obviously have no immediate bearing on the consistency topic of this paper. However, a link table can form a vital part of an assistive IT environment for the working lexicographer whose virtual workbench might include a software tool to help him safeguard reference consistency. Such a reference management tool is currently under development for *elexiko*; this section will present some of its functionality in the light of the preceding remarks.

In what follows, let *D* be the XML document of the dictionary entry currently being edited. In the most basic case, work on the article is done in a generic XML editor. Without a reference management tool, editing references in *elexiko* looks as follows:

- The lexicographer inserts a sense-related item in the entry *D*.
- In the online version of *elexiko*, the lexicographer has to check which senses and subsenses constitute the correct reference target.
- The corresponding IDs of the reference target (lemma/sense/subsense) must be looked up in the *elexiko*-database and manually copied into the entry *D*.
- After completing the entry, the lexicographer has to check the consistency of sense-related items in D in correspondence to the ones in the target entries; this procedure has to be done in the online version.

A reference management tool as a separate application facilitates lexicographic work in a significant way: When entry *D* is opened in the XML editor, the tool enumerates all paradigmatic references in other articles to word senses in *D* (incoming references) as well as all paradigmatic references in *D* to other articles (outgoing references). For each incoming reference in the list, the management tool displays current status information regarding to what extent the consistency criteria given in Section 3 are met for unidirectional as well as bidirectional references. Where an incoming reference is not yet complete because the source article was compiled before editing *D* so that the appropriate target word sense IDs are missing, authors can update the source document with only a few mouse clicks just by choosing from a list of all the word senses in *D*.

In a similar vein, the management tool automatically checks whether all currently outgoing references are correctly specified. The lexicographer can select any of these references and let the program fill in missing details on the desired target word sense by simply choosing from a list. Additionally, the table of outgoing references can be used to speed up navigation within *D* in the editor. A sample screenshot of the management tool developed for *elexiko* is shown in Figure 3 (see below).

Apart from securing consistency with respect to references from and to individual dictionary entries, a reference management tool should also provide tools to scan an entire lexicographic database for

- inconsistent (incorrectly specified) references, in particular references pointing to inexistent entries or word senses within entries; or
- missing reverse references for unidirectional references of an obligatorily bidirectional type.

In *elexiko*, article editing is done in a standard XML editor with a Java API that is used by the reference management application for obtaining the current contents of the active document, navigating within the document, inserting data into it, and so on. On the other hand, the reference manager communicates with the Oracle database system using a standard JDBC interface. The management tool parses the active XML editor document in order to obtain a list of outgoing references. For incoming references, the link table of the database system is used.

7. Conclusion and Prospects

In this paper, we have presented a robust, conceptually parsimonious, and linguistically sound solution to handle cross-references between sense-related entries in an electronic dictionary. We have argued that in typical cases, modelling cross-references with separate data structures simply shifts the sources of possible inconsistencies to another place and merely introduces additional conceptual complexity. Therefore, we suggest to keep information on cross-references strictly local to the respective source entries. To enhance performance of database retrieval, information related to cross-references is additionally kept in a separate, relationally stored link table that is automatically updated whenever entries are altered or added. Taking advantage of such a table, reference management software can then continually screen for referential conflicts while a dictionary entry is being edited and easily check the overall referential consistency of a dictionary database.

Our approach is well suited to a setting where several independent dictionaries are to be gradually integrated into a global database environment with cross-dictionary references. It can easily be extended to other kinds of cross-references between and even within dictionary entries.

The task of visualising lexicographic reference structure is a lucid example of the practical use to which our approach can be put. Figures 4.1 and 4.2 are based on the output of a visualisation tool developed for *elexiko*. Figure 4.1 shows the paradigmatic relations given in the entry for the three word (sub)senses of the entry *Familie* as a directed graph. The program is able to traverse long chains of cross-references from one word sense to the next. In this way, graphs with several thousands nodes (word senses) can be constructed recursively. Calculating such huge graphs on the basis of parsing dictionary entries alone would hardly be feasible; with the use of a link table, it becomes a matter of seconds. In Figure 4.2, some incoming references for the word sense “relatives” are displayed with a recursion depth of 2.

Such a visualisation of paradigmatic structures may be useful for lexicographers for checking a longer chain of

paradigmatically associated entries as well as for navigational tasks provided for dictionary users.

To sum up, our proposal is founded on a fine-grained division of labour: On the one hand, lexicographic reference information that is specific and relevant to an individual entry is represented in the entry itself; on the other hand, further facts about sense relations, such as chains of ever more specific hyponyms of a word sense, can then be inferred efficiently through the use of a link table. This link table not only allows for fast and comfortable consistency checking routines but also for more flexible ways to make use of reference information in an electronic dictionary.

Figure 3. GUI of the reference management software for *elexiko*.

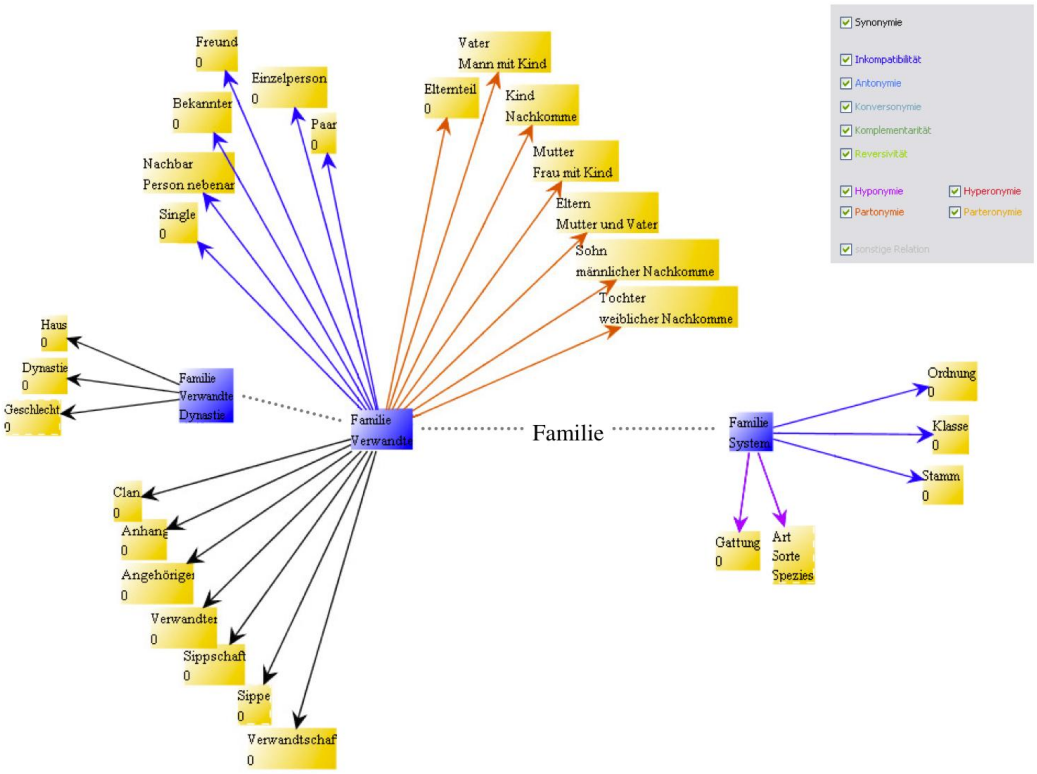

Figure 4.1. Visualization of outgoing references of the *elexiko* entry *Familie,* sense "Verwandte" („relatives") (recursion depth of 1). The boxes represent word (sub)senses and indicate lemma sign, sense ID, and, where applicable, subsense ID. Arrows stand for unidirectional paradigmatic (sense) relations whose type is marked by colour.

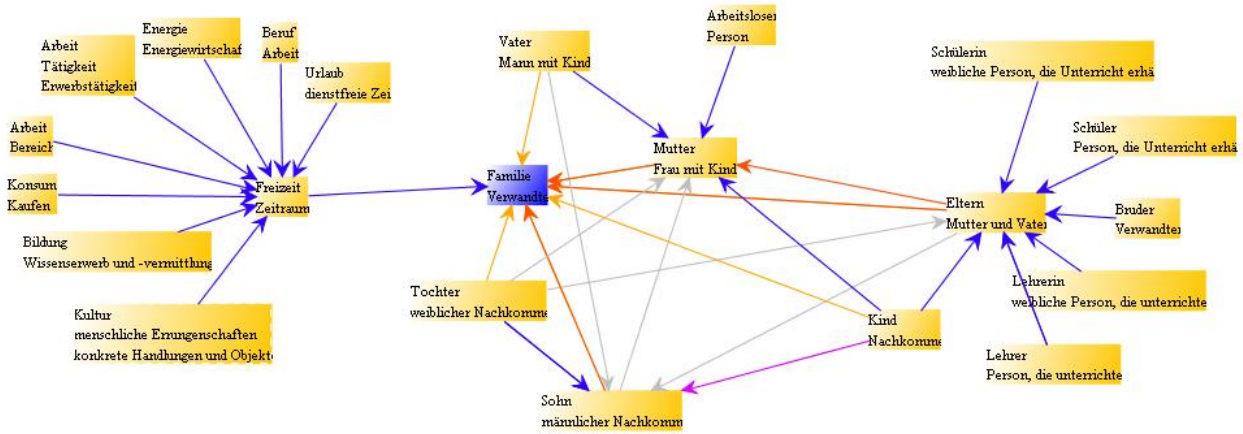

Figure 4.2. Visualization of incoming references of the *elexiko* entry *Familie*, sense "Verwandte" („relatives") (recursion depth of 2). The boxes represent word (sub)senses and indicate lemma sign, sense ID, and, where applicable, subsense ID. Arrows stand for unidirectional paradigmatic (sense) relations whose type is marked by colour.

8. References

Bosch, P. (1979). Synonymie im Kontext. Ein Nachwort. In W.V.O. Quine, *Von einem logischen Standpunkt*. Berlin: Ullstein, pp. 161--172.

Duden (2007). *Das Synonymwörterbuch.* 4th edition, Mannheim/Leipzig/Wien/Zürich: Dudenverlag.

elexiko (2003 seqq.). In *OWID – Online Wortschatz-Informationssystem Deutsch*, Mannheim: Institut für Deutsche Sprache, www.owid.de/elexiko_/index.html

Haß, U. (Ed.) (2005). Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz. Schriften des Instituts für Deutsche Sprache. Berlin, New York: de Gruyter.

Hausmann, F.J., Wiegand, H.E. (1989). Component Parts and Structures of General Monolingual Dictionaries: A Survey. In F.J. Hausmann et al. (Eds.), Wörterbücher / Dictionaries / Dictionnaires. An International Encyclopedia of Lexicography. Berlin, New York: de Gruyter, pp. 328--360.

Klosa, A., Schnörch, U., Storjohann, P. (2006). ELEXIKO - A lexical and lexicological, corpus-based hypertext information system at the Institut für Deusche Sprache, Mannheim. In E. Corino, C. Marello, C. Onesti (Eds.), *Atti del XII Congresso Internazionale di Lessicografia. Torino, 6-9 settembre 2006* (Proceedings of the 12th EURALEX International Congress). Vol. 1. Alessandria: Edizioni dell'Orso, pp. 425--430.

Lew, R. (2007). Linguistic semantics and lexicography: A troubled relationship. In M. Fabiszak (Ed.), *Language and Meaning. Cognitive and Functional Perspectives*. Frankfurt a.M.: Peter Lang, pp. 217--224.

Müller-Spitzer, C. (2007). Vernetzungsstrukturen lexikografischer Daten und ihre XML-basierte Modellierung. *Hermes* 38, pp. 137--171.

Müller-Spitzer, C. (2010). The Consistency of Sense-Related Items in Dictionaries. Current Status, Proposals for Modelling and Potential Applications in Lexicographic Practice. In P. Storjohann (Ed.), *Lexical-semantic relations from theoretical and practical perspectives*. Lingvisticæ Investigationes Supplementa. Amsterdam/New York: Benjamins (forthcoming).

Müller-Spitzer, C., Schneider, R. (2009). Ein XML-basiertes Datenbanksystem für digitale Wörterbücher – Ein Werkstattbericht aus dem Institut für Deutsche Sprache. *it-Information Technology* 51(4), pp. 197--206.

Storjohann, P. (2006). Kontextuelle Variabilität synonymer Relationen. OPAL – Online publizierte Arbeiten zur Linguistik 2006(1), Mannheim: Institut für Deutsche Sprache.

Storjohann, P. (2009). Plesionymy: A case of synonymy or contrast? *Journal of Pragmatics* 41(11), pp. 2140--2158.

Storjohann, P. (2010). Colligational patterns in a corpus and their lexicographic documentation. In M. Mahlberg, V. González-Díaz, & C. Smith (Eds.), *Proceedings of the Corpus Linguistics Conference 2009 in Liverpool*. (published online under: http://ucrel.lancs.ac.uk/publications/CL2009/)